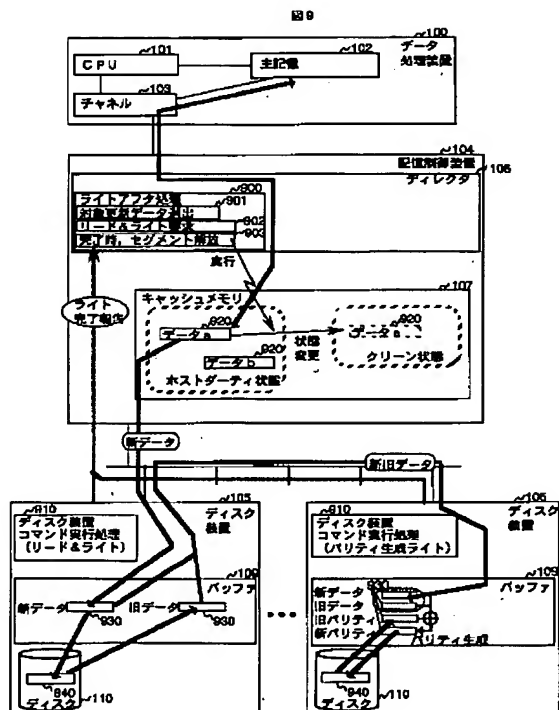


(11)特許出願公開番号

(43)公開日 平成10年(1998)11月24日

F I		
G 0 6 F	3/06	3 0 5 C
		5 4 0
	12/16	3 2 0 L
G 1 1 B	20/18	5 7 0 Z



【特許請求の範囲】

【請求項 1】それぞれのパリティグループが m 個 ($m \geq 1$) のデータと前記データを回復するための冗長データから構成され、少なくとも 1 つ以上の前記パリティグループに属する 1 つ以上の前記データ、あるいは、1 つ以上の前記冗長データを格納する記憶媒体を有する 1 つ以上の記憶装置と、データ処理装置と、前記データ処理装置と前記記憶装置との間に介在し、前記データ処理装置と前記記憶装置とのデータ転送を制御する記憶制御装置からなる記憶装置サブシステムにおいて、前記記憶制御装置が、前記パリティグループに属する一つ以上の前記データの更新値を前記データを格納する一つ以上の第 1 の記憶装置に送信する手段を有し、前記第 1 の記憶装置が前記記憶制御装置から前記データの更新値を受信する手段と、前記データの更新前の値を前記記憶媒体から読み出す手段と、前記データの更新値を前記記憶媒体に書き込む手段と、前記データの更新値と更新前の値を前記冗長データを格納する第 2 の記憶装置に直接送信する手段を有し、前記第 2 の記憶装置が前記第 1 の記憶装置から前記データの更新値と更新前の値を受信する手段と、前記冗長データの更新前の値を前記記憶媒体から読み出す手段と、前記データの更新値と更新前の値と前記冗長データの更新前の値から前記冗長データの更新値を作成する手段と、前記冗長データの更新値を前記記憶媒体に書き込む手段を有することを特徴とする記憶装置サブシステム。

【請求項 2】それぞれのパリティグループが m 個 ($m \geq 1$) のデータと前記データを回復するための冗長データから構成され、少なくとも 1 つ以上の前記パリティグループに属する 1 つ以上の前記データ、あるいは、1 つ以上の前記冗長データを格納する記憶媒体を有する 1 つ以上の記憶装置と、データ処理装置と、前記データ処理装置と前記記憶装置との間に介在し、前記データ処理装置と前記記憶装置とのデータ転送を制御する記憶制御装置からなる記憶装置サブシステムにおいて、前記記憶制御装置が、前記パリティグループに属する一つ以上の前記データの更新値を、前記データを格納する 1 つ以上の第 1 の記憶装置に送信する手段と、前記データの更新前の値を前記第 1 の記憶装置から受信する手段と、前記データの更新値と更新前の値を前記冗長データを格納する第 2 の記憶装置に送信する手段を有し、前記第 1 の記憶装置が前記記憶制御装置から前記データの更新値を受信する手段と、前記データの更新前の値を前記記憶媒体から読み出す手段と、前記データの更新値を前記記憶媒体に書き込む手段と、前記データの更新前の値を前記記憶制御装置に送信する手段を有し、前記第 2 の記憶装置が前記記憶制御装置から前記データの更新値と更新前の値を受信する手段と、前記冗長データの更新前の値を前記記憶媒体から読み出す手段と、前記データの更新値と更新前の値と前記冗長データの更新前の値から前記冗長デー

タの更新値を作成する手段と、前記冗長データの更新値を前記記憶媒体に書き込む手段を有することを特徴とする記憶装置サブシステム。

【請求項 3】請求項 2 で、前記記憶制御装置が、複数の前記データの更新値と更新前の値を同時に格納できるキャッシュメモリを有し、前記第 1 の記憶装置から前記データの更新前の値を受信後、前記データの更新値とともに前記キャッシュメモリに溜め置き、前記データの更新前の記憶装置への反映処理とは非同期に、シーク／サーチ動作時間を考慮したスケジュールにより、前記第 2 の記憶装置上の前記冗長データへの反映処理を行うことを特徴とする記憶装置サブシステム。

【請求項 4】請求項 1 で、前記第 2 の記憶装置が、複数の前記データの更新値と更新前の値を同時に格納できるバッファメモリを有し、前記第 1 の記憶装置から前記データの更新値と更新前の値を受信後、バッファメモリに溜め置き、前記データの更新前の記憶装置への反映処理とは非同期に、シーク／サーチ動作時間を考慮したスケジュールにより、前記第 2 の記憶装置上の前記冗長データへの反映処理を行うことを特徴とする記憶装置サブシステム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、ディスクアレイ向きの高機能ディスク装置、および、ディスクアレイ向きの高機能ディスク装置と記憶制御装置により構成される記憶装置サブシステムに関する。

【0002】

【従来の技術】発明に最も近い公知例として、シカゴのイリノイ大学で開かれた ACMS SIGMOD 会議において発表された D. Patterson, G. Gibson, and R. H. Kartz; A Case for Redundant Array of Inexpensive Disks (RAID), ACM SIGMOD Conference, Chicago, IL, (June 1988), pp. 109-116 なる論文が知られている。

【0003】Patterson の論文は、ディスクアレイ上のデータ配置に関する技術を開示したものである。ディスクアレイシステムでは、物理的には複数のディスク装置をデータ処理装置に対しては 1 台の論理的なディスク装置に見せかけることで高性能化を実現している。また、データを格納したディスク装置に障害が発生した場合、障害が発生したディスク装置に記憶されているデータの回復を行うことを可能とする冗長データをデータを記憶したディスク装置とは別のディスク装置に格納しておくことにより高信頼化を達成している。

【0004】通常、ディスク装置に格納されたデータのリード／ライト単位をレコードと呼ぶ。ただし、ディスクアレイの場合、データ処理装置から見たリードライト単位であるレコードと、ディスク装置に実際に記録され

3

るレコードとではデータ長が異なる場合がある。以下、前者を論理レコード、後者を物理レコードと呼ぶ。

【0005】Pattersonの論文では、いくつかの論理レコードの配置方法がRAID1からRAID5の5段階のレベルのRAID構成として紹介され、性能および信頼性の検討結果が報告されている。このうち、RAID5のディスクアレイは各論理レコードを1つの物理レコードとしてディスク装置に格納し、互いに異なるディスク装置に格納された m ($m \geq 1$) 個の論理レコードに対して、その排他的論理和であるパリティと呼ばれる冗長データを作成し、1つの物理レコードとしてディスク装置に格納したものである。以下、処理装置が直接リード/ライトする論理レコードを格納した物理レコードをデータレコード、パリティを格納した物理レコードをパリティレコード、データレコードを記憶するディスク装置をデータディスク装置、パリティレコードを記憶するディスク装置をパリティディスク装置と呼ぶ。また、 m 個のデータレコードとパリティレコードから構成されるレコードのグループをパリティグループと呼ぶ。RAID5のディスクアレイは、更新のあったデータレコードに対して、パリティ生成時にデータレコードとパリティレコードの更新前の値である旧データと旧パリティをそれぞれのレコードが格納されるディスク装置からリードし、データレコードの更新値である新データと排他的論理和をとってパリティレコードの更新値である新パリティを生成し、ディスク装置へ新パリティをライトする必要があるため、ライト処理に付随するオーバーヘッドは大きい。このようなパリティ生成処理を効率よく行い、システムの高性能化を実現する技術としては次のような技術がある。

【0006】特開平4-245342号に開示される技術では、論理レコード(=データレコード)に対する更新処理をキャッシュ上で実行し、新パリティの作成や新データと新パリティのディスク装置への書き込みは後から実行する。ある更新データレコードに対して新パリティの作成を開始する前に、同じパリティレコードに対応する他のデータレコードに更新があった場合、新パリティの作成をまとめて実行することにより高性能化を実現している。このように、処理装置からの論理レコード更新に対し、ディスク装置への反映を非同期で行う方式を一般にライトアフトと呼ぶ。

【0007】特開平5-46324号には、新パリティの生成に必要な情報をパリティレコードのあるパリティディスク装置に転送し、ディスク装置側で新パリティを生成することで、記憶制御装置の負荷や、データ更新時に発生する記憶制御装置とディスク装置との間のデータ転送回数を削減し、ディスクアレイシステムの高性能化を図る技術が開示されている。この技術では、パリティ生成に必要な情報をパリティディスク装置に送出する方法の一つとして、記憶制御装置と各ディスク装置をプロ

4

ードキャスト可能なネットワークで接続し、パリティグループの1データの新データをデータディスク装置とパリティディスク装置にブロードキャストして、旧データはデータディスク装置からパリティディスク装置へ直接転送することで、インタフェースにおける転送量を低減する方法が開示されている。

【0008】また、転送量の低減を図る技術としては、特開平7-261946号で、更新データレコードの新データをデータディスク装置に転送し、ディスク装置側で旧データを読んで新データを書き込み、新旧データから中間値を生成し、生成した中間値をパリティディスク装置に直接転送する技術が開示されている。新旧データをそのまま転送するのではなく、中間値に変換して転送することで、データディスク装置とパリティディスク装置間のデータ転送量を低減できる。

【0009】

【発明が解決しようとする課題】RAID5のディスクアレイを用いた場合、記憶制御装置がデータ処理装置からライト要求を受けて、データレコードを更新するのに伴って、新パリティの生成、およびパリティレコードの更新が必要とする。 m 個のデータレコードとパリティレコードから構成されるパリティグループの新パリティ生成方法としては、(a)パリティグループの m 個のデータレコードすべての排他的論理和により算出する方法の他に、排他的論理和の性質を利用して、(b)パリティグループのいくつかの更新データレコードの旧データと新データとパリティレコードの旧パリティとの排他的論理和から算出する方法がある。通常、(b)の方法の方が、新パリティの生成に必要な情報を揃えるためのオーバーヘッドが小さいため、(b)の方法を前提に以下の説明を行う。パリティ生成に必要な情報を揃えるための転送処理として、(b)に示した値を得る処理を実行すると、パリティグループに更新データレコードが一つだけの場合でも、更新データレコードの旧データ、および対応する旧パリティをディスク装置から読み込むために2回、新データと新パリティをディスク装置に書き込むために2回、計4回の記憶制御装置とディスク装置間のデータ転送が生じ、キャッシュメモリの割り当てとディスク装置間のバス確保などのオーバーヘッドが発生する。さらに、そろえた情報を元に新パリティを生成するオーバーヘッドも生じる。

【0010】特開平5-46324号に開示された技術は、ディスク装置に新パリティを作成する機能を持たせることで、記憶制御装置の負荷をある程度分散させることができる。また、データ転送は、更新データレコードの新データをデータディスク装置とパリティディスク装置にブロードキャストするのに1回、更新データレコードの旧データを読み込んでパリティディスク装置に転送するのに1回、計2回で済むことになる。

【0011】ただ、この技術では、ブロードキャストさ

れた新データをパリティディスク装置で受信しても、データディスク装置から対応する旧データを受信するまでは新パリティ生成を行なえない。このため、複数の更新データレコードの新データをディスク装置内のバッファメモリに保持することになるが、保持した複数の新データのスケジュールにはほとんど自由度がなく、あくまで旧データ到着順とせざるを得ない。仮に、平均シーク時間を短縮するべく、現ディスクヘッド位置から最短シークとなる新データから処理しようとするなら、旧データが未到着で処理可能でない新データとともに、旧データが到着して処理可能となった新データをも複数保持せねばならず、ディスク装置に搭載するバッファメモリを大容量のものにする必要が生じ、製造コストを押し上げることになる。

【0012】一方、特開平7-261946号公報の技術では、データディスク装置に新旧データから中間値を作成する手段を設けることで、データ転送を、データディスク装置への新データ転送に1回、パリティディスク装置への新旧データの中間値転送に1回、計2回に抑えている。

【0013】しかし、この技術では、新データと旧データの排他的論理和演算をデータディスク装置で、その結果生成した中間値と旧パリティの排他的論理和演算をパリティディスク装置で、それぞれ分割して実行するため、パリティディスク装置で三者の排他的論理和演算を一括実行する方法に比べて、各データを一時的に格納するバッファメモリへのロード／ストア実行回数が増加し、その結果、演算に要する総オーバーヘッドは増大するものと考えられる。

【0014】また、この技術では、データレコードへの更新に伴ってパリティレコードへの更新が発生し、互いに独立に動作しているデータディスク装置とパリティディスク装置へ同じ契機にアクセスを行なうことになる。このため、両ディスク装置の平均シーク時間の低減に対しては何ら考慮されていない。

【0015】本発明の目的は、パリティ生成に要するオーバーヘッドの軽減を図り、記憶装置サブシステム全体のスループット増大を図ることである。

【0016】本発明の別の目的は、パリティ生成におけるデータレコードとパリティレコードの更新を非同期化し、各ディスク装置アクセスに伴う平均シーク時間を軽減し、スループット増大を図ることである。

【0017】

【課題を解決するための手段】本発明では、記憶装置サブシステム全体のスループット増大を図るという目的を実現するために、以下の2つの観点で示した記憶制御方法を実現する記憶装置サブシステムを提供する。

【0018】第一の観点では、記憶制御装置はデータ処理装置から更新され、キャッシュ上に蓄積したデータレコードのライトアフタ処理を行うために、同一パリティ

グループのデータレコードのうち、キャッシュ上に存在するものの新データを対応するデータディスク装置に送信する。データディスク装置では、受信した新データに対する旧データをリードした後、新データをディスクにライトし、新旧データを対応するパリティディスク装置へ転送する。一方、パリティディスク装置では、受信した新旧データに対応するパリティレコードの旧パリティをリードし、新パリティを生成してディスクにライトする。

10 【0019】第二の観点でも、記憶制御装置は、第一の観点と同様に、同一パリティグループの複数のデータレコードの新データを各データディスク装置に転送し、データディスク装置では、対応する旧データをリード後、新データをライトする。第二の観点と異なるのは、ここでリードした旧データを一旦記憶制御装置に転送する点である。記憶制御装置では、ディスク装置への反映が完了し、旧データを読み込んだデータレコードをしばらくキャッシュ上に溜め置き、データディスク装置への反映とは非同期にパリティディスク装置への反映処理を実行する。

20 【0020】これらの結果、パリティ生成のための排他的論理和演算はパリティディスク装置で一括して実行するので、演算の際のメモリロード／ストア回数を削減でき、オーバーヘッドを軽減することができる。さらに、パリティレコードの更新をデータレコードの更新と非同期に実行可能となり、シーク時間を考慮したディスク装置の割り当てスケジューリングを行なうことで、各ディスク装置アクセス時の平均シーク時間を短縮することができる。

30 【0021】以上の内容を適用すると、更新データのライトの際に、ディスク装置でのパリティ生成オーバーヘッド削減、かつ平均シーク時間短縮によるディスク装置占有時間削減を実現することで、記憶装置サブシステム全体のスループットを増大させることが可能となる。

【0022】

【発明の実施の形態】以下、本発明の実施例について説明する。なお、これにより本発明が限定されるものではない。

40 【0023】本発明の実施例は、第1の実施例と第2の実施例がある。

【0024】第1の実施例は、記憶制御装置から新データを受信したデータディスク装置で、旧データを記憶媒体から読み出し、受信した新データを記憶媒体に書き込み、新旧データを併せて、対応するパリティディスク装置に直接送信するものである。

50 【0025】第2の実施例は、記憶制御装置から新データを受信したデータディスク装置で、旧データを記憶媒体から読み出し、受信した新データを記憶媒体に書き込み、旧データを記憶制御装置に送信し、記憶制御装置からパリティディスク装置へ新旧データを送信するもので

ある。

【0026】まず、図1～図11を参照して、第1の実施例を説明する。

【0027】図1は、本発明の対象となる記憶装置サブシステムのブロック図である。

【0028】記憶装置サブシステムは、データ処理装置100と、記憶制御装置104と、1台以上のディスク装置105を接続してなっている。

【0029】データ処理装置100は、CPU101と、主記憶102と、チャンネル103とを有している。 10

【0030】記憶制御装置104は、1つ以上のディレクタ106と、キャッシュメモリ107と、ディレクトリ108と、不揮発性メモリ109と、不揮発性メモリ管理情報110を有している。

【0031】ディレクタ106は、チャンネル103とディスク装置105の間のデータ転送、チャンネル103とキャッシュメモリ107の間のデータ転送およびキャッシュメモリ107とディスク装置105の間のデータ転送を行なう。

【0032】キャッシュメモリ107には、ディスク装置105の中のアクセス頻度の高いデータをロードしておく。このロード処理は、ディレクタ106が実行する。ロードデータの具体的な例は、CPU101からアクセス対象となったデータレコードや、このデータレコードと物理ディスク装置105のディスク位置が近いデータレコードのデータ等である。また根、キャッシュメモリ107には、CPU101から更新がかけられ、ライトアプタで処理する新データを格納する。これらの新データは、キャッシュメモリ107に格納した段階でライト完了報告を出し、更新とは非同期にパリティ生成およびディスク装置105への書き込みを行う。 20

【0033】ディレクトリ108は、キャッシュメモリ107の管理情報を格納する。

【0034】不揮発性メモリ109は、不揮発の媒体であり、キャッシュメモリ107に格納した新データが物理ディスク装置105への反映が完了する前に消失しないよう、新データを格納し、キャッシュメモリ107と併せて二重管理する。

【0035】不揮発性メモリ管理情報110も不揮発の媒体であり、不揮発性メモリ109の管理情報を格納する。 40

【0036】ディスク装置105は、データレコードを記録するディスク110とデータを一時的に格納するバッファ109と記録されたデータを読み書きする装置から構成される。データを読み書きする装置は、排他的論理和演算により新パリティを生成する機能を有する。バッファ113は、複数のデータを格納することができる容量を持ち、必要があれば不揮発の媒体を用いる。

【0037】各ディスク装置105間には互いに接続され、複数のディスク装置105間でのデータの送受信が 50

可能である。

【0038】図2は、本実施例におけるキャッシュメモリ107の構成である。

【0039】キャッシュメモリ107は複数のセグメント200で構成される。一つのセグメント200には、1つのデータレコードもしくはパリティレコードを格納しうる。どのセグメント200にどのレコードを割り当てるかについては、従来公知の方法を用いることができる。レコードが割り当てられていない空き状態のセグメント200も存在しうる。

【0040】図3は、ディレクトリ108の構造である。

【0041】ディレクトリ108は、キャッシュ割り当て情報300、セグメント管理情報301、空きセグメントキューポインタ302から構成される。なお、セグメント管理情報301は、セグメント200の数だけ存在する。ただし、セグメント200に静的に1対1対応するわけではない。以下、各情報を詳細に説明する。

【0042】図4は、キャッシュ割り当て情報300の構造である。データレコードやパリティレコードにセグメント200が割り当てられ、キャッシュメモリ107上に格納されているかを表わす。レコード割り当てポインタ400は、データレコードとパリティレコードに対応して設ける情報である。対応するデータレコードやパリティレコードがキャッシュ107上に格納されていれば、レコード割り当てポインタ400は当該セグメント管理情報301をポイントする。格納されていなければヌル値となる。

【0043】図5は、セグメント管理情報301の中で本発明に関係した情報を表わしたものである。ライト面セグメントポインタ500について以下説明する。データレコードのセグメント管理情報301の場合、新パリティを生成していない新データを格納したセグメント200をポイントする。パリティレコードのセグメント管理情報301の場合、キャッシュメモリ107上で新パリティを生成する際に、一時的に利用するセグメント200をポイントする。以後、ライト面セグメントポインタ501でポイントするこれらのセグメント200をライト面セグメントと呼ぶ。 30

【0044】リード面セグメントポインタ501について以下説明する。データレコードのセグメント管理情報301の場合、新パリティを作成済みの値を格納したセグメント200をポイントする。パリティレコードのセグメント管理情報301の場合、原則的に、本ポインタが示すセグメント200にその値を格納する。以後、リード面セグメントポインタ501でポイントするこれらのセグメント200をリード面セグメントと呼ぶ。

【0045】ホストダーティフラグ502は、ライト面セグメントの値がデータ処理装置100から書き込まれたままで、パリティ生成もディスク装置105への反映

も行なわれていないことを示す。物理ダーティフラグ 502 は、リード面セグメント値がパリティ生成が完了したが、まだディスク装置 105 に書き込まれていないことを示す。パリティ未反映フラグ 502 は、ライト面セグメントの値が、ディスク装置 105 へ反映されたが、パリティ生成がまだ行なわれていないことを示す。ライト面バリッドフラグ 503 はライト面セグメントに、リード面バリッドフラグ 504 はリード面セグメントに有効な値が入っていることを示す。使用中フラグ 505 は、そのセグメント管理情報 301 が割り当てられているデータまたはパリティレコードが何らかの処理で使用中であることを示し、処理間の排他に用いられる。レコード状態 506 は、個々のデータまたはパリティレコードの状態を示す。状態としては、「正常」の他に、ディスク装置 105 の障害等が原因で書き込みが出来ないことを示す「ライト不可」、読み込みが出来ないことを示す「リード不可」などがあり、複数の状態が同時に設定されることもある。空きポインタ 507 は、空き状態にあるセグメント管理情報 301 同士を結合する。

【0046】図 6 は、空きセグメントキューの構造である。図に示すように、空き状態にあるセグメント 200 の管理情報 301 を空きセグメントキューポインタ 302 に結合する。

【0047】本実施例におけるキャッシュメモリ管理方法の説明の最後に、各セグメント 200 に格納されるデータまたはパリティレコードの状態について述べる。

【0048】図 7 は、セグメント 200 に格納されるデータまたはパリティレコードの状態遷移の様子である。処理装置 100 から更新がかかり、パリティ生成を行う前で、ライト面セグメントに新データを格納し、ライト面バリッドフラグ 503 とホストダーティフラグ 502 がオンの状態をホストダーティ状態 700 と呼ぶ。ホストダーティ状態 700 から、当該データレコードに対するパリティ生成が完了し、新データを格納したライト面セグメントをリード面セグメントに切り替え、リード面バリッドフラグ 504 と物理ダーティフラグをオンに、ホストダーティフラグ 502 をオフにした状態を物理ダーティ状態 701 と呼ぶ。キャッシュ 107 上で生成した新パリティも物理ダーティ状態 701 で保持する。また、ホストダーティ状態から、当該データレコードのディスク装置 105 より旧データを読んでリード面セグメントに格納し、新データをディスク装置 105 へ反映し、ホストダーティフラグ 502 をオフ、パリティ未反映フラグをオンにした状態をパリティ未反映状態 702 とする。さらに、物理ダーティ状態 701 のセグメント 200 に対して、当該データまたはパリティレコードのディスク装置 105 への反映が完了し、物理ダーティフラグ 502 をオフにした状態、および、パリティ未反映状態 702 のセグメント 200 に対して、当該データレコードのパリティ生成が完了し、パリティ未反映フラグ

502 をオフにした状態をクリーン状態 703 と呼ぶ。

【0049】以下の説明では、あるセグメント 200 をホストダーティ状態、物理ダーティ状態、パリティ未反映状態、クリーン状態にすることを、単にホストダーティ化、物理ダーティ化、パリティ未反映化、クリーン化と呼ぶ。また、あるデータ又はパリティセグメント 200 の割り当てを解除して、セグメント 200 およびセグメント管理情報 301 を空き状態にすることを、セグメント破棄と呼ぶ。セグメント 200 破棄の際には、設定されているフラグ、状態、ポインタを全てクリアする。

【0050】図 8 は、本実施例におけるディスク装置内バッファ 109 の構成である。バッファ 109 は複数のバッファセグメント 800 で構成される。一つのバッファセグメント 800 には 1 物理レコード分の値を格納でき、ディスク装置 105 と記憶制御装置 104 間でやりとりするデータまたはパリティレコードの値を格納する。リード処理時はディスク 110 からリードされた値が一時格納され、ライト処理時には記憶制御装置 104 から転送された書き込み値が一時格納される。

【0051】次に本実施例での記憶制御装置 104 とディスク装置 105 の動作を説明する。

【0052】図 9 は、本実施例における記憶制御装置 104 とディスク装置 105 の動作を詳細に表わした図である。

【0053】まず、データ処理装置 100 によりライトされたデータレコード 920 に対してセグメント 200 を割り当て、新データを格納する。この段階で、ライト処理の完了をデータ処理装置 100 に報告し、ホストダーティ状態の当該データレコード 920 をライト要求とは非同期にディスク装置 105 に反映する。

【0054】この第 1 の実施例では、キャッシュ 107 上に保持されているホストダーティ状態のデータレコード 920 の中から、ライトアプタ処理対象となるものを選択し、対象データレコード 920 の新データを、当該データを格納するデータディスク装置 105 に送出する。次に、データディスク装置 105 では、受信した新データに対し、旧データのディスク 110 からの読みだしと新データのディスク 110 への書き込み、および対応するパリティを格納するパリティディスク装置 105 への新旧データの送出を行なう。以下、データディスク装置 105 でのこれら一連の処理をリード&ライト処理と呼ぶ。新旧データを受信したパリティディスク装置 105 では、旧パリティのディスク 110 からの読み出しと新パリティ生成、および新パリティのディスク 110 への書き込みを行なう。以下、パリティディスク装置 105 でのこれら一連の処理をパリティ生成ライト処理と呼ぶ。最後に、要求したデータディスク装置からの処理完了報告を受けたライト転送処理 900 が、当該更新データレコード 920 のセグメント 200 をクリーン化、およびパリティレコードのセグメント 200 を破棄す

る。

【0055】以下では、ライトアフト処理900と、ディスク装置コマンド実行処理910について順に説明する。

【0056】第一に、ライトアフト処理900について説明する。

【0057】記憶制御装置104からディスク装置105への処理要求は、処理指定情報と処理対象データから構成される。指定情報は以下のような内容である。

【0058】指定情報1…要求処理種別を指定する。要求処理には、リード、ライト、リード&ライト、パリティ生成ライトなどがある。本実施例において、対象更新データレコードの新データを送信する際には要求処理種別にはリード&ライトを、データディスク装置105からパリティディスク装置105へ新旧データを送信する際にはパリティ生成ライトをそれぞれ設定する。

【0059】指定情報2…ディスク装置での処理対象範囲を指定する。

【0060】処理対象データには、ライト要求やリード&ライト要求の場合には更新データの新データが該当する。

【0061】ライトアフト処理900では、セグメント200に格納されている更新データレコード920のリード&ライト要求をデータディスク装置105に送信し、データレコードとパリティレコードの更新完了を待って、データレコードに割り当てたセグメント200のクリーン化などの処理を行なう。

【0062】図10はライトアフト処理900の処理フロー図である。

【0063】ステップ1000では、セグメント200に格納され、キャッシュ107上に蓄積されたホストデータ状態のデータレコードの中から、処理対象を選び出す。このとき、同一パリティグループに属する複数の更新データレコードが存在する場合は、同時に処理対象とする。

【0064】ステップ1001では、ステップ1000で選出した各更新データレコード920に対応するパリティレコードにセグメント200を割り当て、当該セグメント管理情報301の使用フラグ505をオンにする。こうして、他処理による当該パリティグループのデータに対するパリティ生成を禁じ、排他を実現する。

【0065】ステップ1002では、ステップ1000で選出した各更新データレコード1420のセグメント管理情報301の使用フラグ505をオンにする。

【0066】ステップ1003では、処理対象の更新データレコード920の新データと処理指定情報からなるリード&ライト要求を各データディスク装置105に送出し、データディスク装置105におけるリード&ライト処理、およびパリティディスク装置105におけるパリティ生成ライト処理を行なわせる。

【0067】ステップ1004では、データディスク装置105とパリティディスク装置105に対する更新の完了した更新データレコード920のセグメント200をクリーン化し、使用中フラグ505をオフする。ここで、セグメント200を破棄ではなく、クリーン化するのは、以降のアクセスによる当該データレコードへのキャッシュヒットを期待するためである。

【0068】ステップ1005では、当該更新データレコード920のライト処理のために割り当てたパリティセグメント200を破棄する。

【0069】第二に、ディスク装置コマンド実行処理910について説明する。

【0070】ライト転送処理900により転送されたリード&ライト要求を受信したデータディスク装置105では、当該データレコードの旧データリードと新データライトを実行後、パリティディスク装置105へ新旧データと共にパリティ生成ライト要求を送信する。パリティ生成ライト要求を受信したパリティディスク装置105では、旧パリティリードと新パリティ生成、および新パリティライトを行なう。

【0071】図11はディスク装置コマンド実行処理910の本発明に関係した部分のみを抜粋した処理フロー図である。

【0072】ステップ1100では、受信した処理指定情報の要求処理種別を調べる。処理種別がリード&ライトの場合にはステップ1101に遷移し、パリティ生成ライトの場合にはステップ1106にジャンプする。

【0073】ステップ1101では、処理指定情報とともに送信されてきた対象データレコードの新データをディスク装置105内のバッファセグメント930に取り込む。

【0074】ステップ1102では、指定情報2で指定されたディスク110上の格納位置から、対象データレコードの旧データをリードし、新データを格納したのとは別のバッファセグメント930に格納する。

【0075】ステップ1103では、ディスク110上の格納位置に、バッファセグメント930に格納した対象データレコードの新データをライトする。

【0076】ステップ1104では、各々バッファセグメント930に格納した新旧データと処理指定情報からなるパリティ生成ライト要求をパリティディスク装置105に送出する。

【0077】ステップ1105では、要求処理が完了したことをライトアフト処理900に報告し、処理を終了する。

【0078】ステップ1106では、パリティ生成ライト処理を行なうために、処理指定情報とともに送信されてきた新旧データをバッファセグメント930に各々取り込む。

【0079】このとき、対象データレコードが複数存在

する場合には、データレコード毎にパリティ生成ライトを行なっても良いし、全データレコード分の新旧データを同時にバッファセグメント930に取り込み、新パリティを一括生成しても構わない。また、パリティ生成ライト要求の新旧データを複数のパリティレコード分バッファ内に109にため込み、シーク時間最短のものから対象にするなど、適当なスケジューリングを行なっても構わない。

【0080】ステップ1107では、パリティ生成ライト処理を行なう。まず、ディスク110から新パリティ生成に必要な旧パリティをリードして、バッファセグメント930に格納する。次にバッファ109上の新データと旧データ、および旧パリティから新パリティを生成し、また別のバッファセグメント930に格納する。最後に、ディスク110上の先ほどリードしたパリティ格納位置に、新パリティをライトする。

【0081】次に、図12～図15を参照して、第2の実施例を説明する。

【0082】図12は、第2の実施例における記憶制御装置104とディスク装置105の動作を詳細に表した図である。

【0083】この第2の実施例では、新データのデータディスク装置105への反映が終った段階で、データレコード920をパリティ未反映化して、キャッシュ上に保持し、パリティディスク装置105への反映処理を別の契機に行なう。まず、ホストダーティ状態のデータレコード920の中から、データ反映処理対象となるものを選択し、当該データレコード920の新データを、当該データレコードを格納するデータディスク装置105に送出する。次に、データディスク装置105では、受信した新データに対し、旧データのディスク110からの読みだしと新データのディスク110への書き込みを行ない、記憶制御装置104に旧データを送出する。以下、データディスク装置105での、これら一連の処理も第1の実施例と同様に、リード&ライト処理と呼ぶ。記憶制御装置104では、旧データを受信したデータレコード920をパリティ未反映化して、当該データレコード920に対する処理を一旦打ち切る。キャッシュ107上に保留されたデータレコード920に対しては、この後、別の契機でパリティ反映処理を実行し、完了後、セグメント200をクリーン化する。

【0084】以下では、データ反映処理1200と、パリティ反映処理1210と、ディスク装置コマンド実行処理1220について順に説明する。

【0085】第一に、データ反映処理1200について説明する。

【0086】データ反映処理1200では、更新データレコード920のリード&ライト要求をデータディスク装置105に送信し、データレコードの更新完了を待つて、受信した旧データをセグメント200に格納し、デ

ータレコード920のパリティ未反映化などの処理を行なう。

【0087】図13はデータ反映処理1200の処理フロー図である。

【0088】ステップ1300では、セグメント200に格納され、キャッシュ107上に蓄積されたホストダーティ状態のデータレコード930の中から、処理対象を選び出す。

【0089】ステップ1301では、ステップ1301で選出した更新データレコード920のセグメント管理情報301の使用フラグ505をオンにする。

【0090】ステップ1302では、処理対象の更新データレコード920の新データと処理指定情報からなるリード&ライト要求をデータディスク装置105に送出し、リード&ライト処理を行なわせる。

【0091】ステップ1303では、データディスク装置105に対する反映の完了した更新データレコード920をパリティ未反映化し、使用中フラグをオフする。

【0092】第二に、パリティ反映処理1210について説明する。

【0093】パリティ反映処理1210では、データ反映処理1200によってデータディスク装置105への反映処理が完了して、キャッシュ107上にパリティ未反映状態で蓄積されているデータレコード920のパリティ生成ライト要求をパリティディスク装置105に送信し、パリティレコードの更新完了を待つて、データレコード920のクリーン化などの処理を行なう。

【0094】図14はパリティ反映処理1210の処理フロー図である。

【0095】ステップ1400では、セグメント200に格納され、キャッシュ107上に蓄積されたパリティ未反映状態のデータレコード930の中から、処理対象を選び出す。

【0096】ステップ1401では、ステップ1401で選出した更新データレコード920のセグメント管理情報301の使用フラグ505をオンにする。

【0097】ステップ1402では、処理対象の更新データレコード920の新旧データと処理指定情報からなるパリティ生成ライト要求をパリティディスク装置105に送出し、パリティ生成ライト処理を行なわせる。

【0098】ステップ1403では、パリティディスク装置105に対する反映の完了した更新データレコード920をクリーン化し、使用中フラグをオフする。

【0099】第三に、ディスク装置コマンド実行処理1220について説明する。

【0100】データ反映処理1200により転送されたリード&ライト要求を受信したデータディスク装置105では、当該データレコードの旧データリードと新データライトを実行後、記憶制御装置104へ旧データを転送する。また、パリティ反映処理1210からパリティ

生成ライト要求を受信したパリティディスク装置105では、旧パリティリードと新パリティ生成、および新パリティライトを行なう。

【0101】図15はディスク装置コマンド実行処理1220の本発明に関係した部分のみを抜粋した処理フロー図である。

【0102】ステップ1500からステップ1507のうち、ステップ1504以外は図11と全く同じであるので、説明は繰り返さない。

【0103】ステップ1504では、ステップ1502でディスク110から読み込み、バッファセグメント930に格納した旧データを記憶制御装置104に送出する。

【0104】

【発明の効果】本発明の記憶制御方法および記憶装置サブシステムによれば、データライト時の記憶制御装置の負荷軽減、かつ平均シーク／サーチ時間短縮を実現したパリティ生成ライトが可能となり、記憶制御装置のオーバヘッド削減の助長やドライブ占有時間削減などの効果により記憶装置サブシステム全体のスループットを増大させることができる。

【図面の簡単な説明】

【図1】本発明が対象とする記憶装置サブシステムのブロック図である。

【図2】本発明の実施例におけるキャッシュメモリの構成例示図である。

【図3】本発明の実施例におけるディレクトリの構成例示図である。

【図4】本発明の実施例におけるキャッシュ割り当て情報の構成例示図である。

【図5】本発明の実施例におけるセグメント管理情報の構成例示図である。

【図6】本発明の実施例における空きセグメントキューの構成例示図である。

【図7】本発明の実施例におけるセグメント状態遷移図である。

【図8】本発明の実施例におけるディスク装置内のバッファの構成例示図である。

【図9】本発明の第1の実施例において、キャッシュ上

に保持された、データ処理装置からの新データをディスク装置へ書き込む際の、制御装置およびディスク装置の動作を示すブロック図である。

【図10】本発明の第1の実施例において、更新データレコードのリード&ライト要求を行うライトアフタ処理の処理フロー図である。

【図11】本発明の第1の実施例において、記憶制御装置からの要求を処理するディスク装置コマンド実行処理の処理フロー図である。

【図12】本発明の第2の実施例において、キャッシュ上に保持された、データ処理装置からの新データをディスク装置へ書き込む際の、制御装置およびディスク装置の動作を示すブロック図である。

【図13】本発明の第2の実施例において、更新データレコードのリード&ライト要求を行うデータ反映処理の処理フロー図である。

【図14】本発明の第2の実施例において、パリティ未反映状態のデータレコードのパリティ生成ライト要求を行うパリティ反映処理の処理フロー図である。

【図15】本発明の第2の実施例において、記憶制御装置からの要求を処理するディスク装置コマンド実行処理の処理フロー図である。

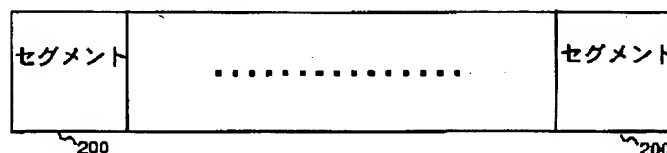
【符号の説明】

100…データ処理装置、 101…CPU、 102…主記憶、 103…チャネル、 104…記憶制御装置、 105…ディスク装置、 106…ディレクタ、 107…キャッシュメモリ、 108…キャッシュディレクトリ、 109…バッファ、 110…ディスク、 111…NVS、 112…NVS管理情報、 200…セグメント、 301…セグメント管理情報、 800…バッファセグメント、 900…第1の実施例のライトアフタ転送処理部、 910…第1の実施例のディスク装置コマンド実行処理部、 920…キャッシュ上のデータセグメント、 930…バッファに格納されたライトアフタ要求、 1200…第2の実施例のデータ反映処理部、 1210…第2の実施例のパリティ反映処理部、 1220…第2の実施例のディスク装置コマンド実行処理部。

【図2】

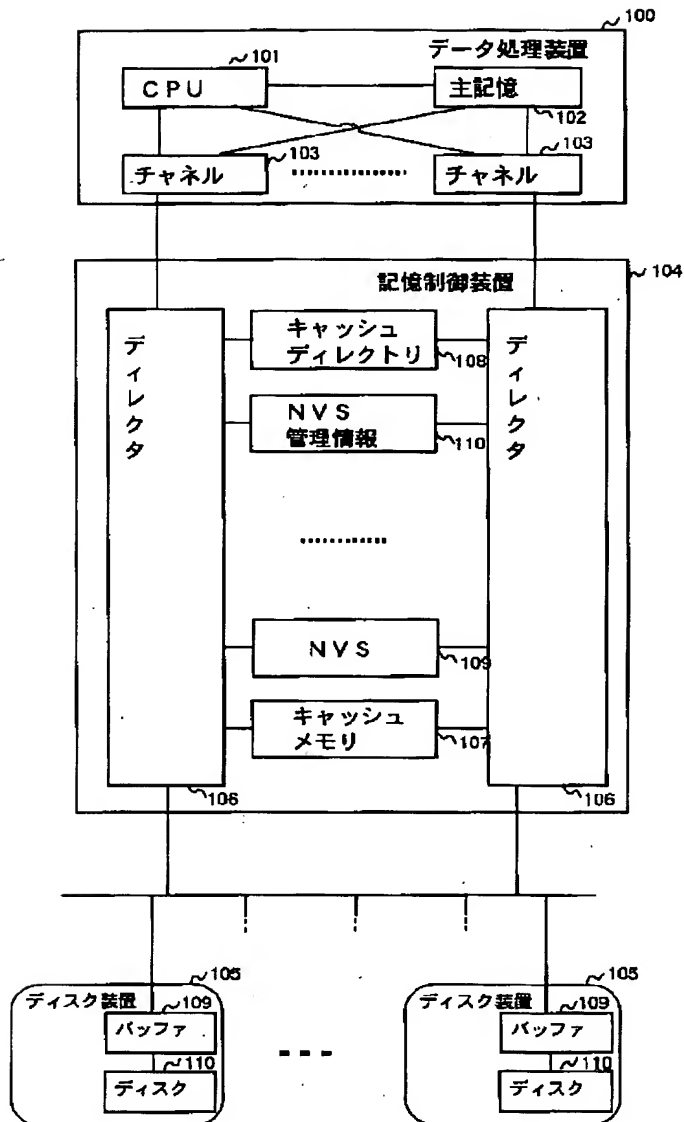
図2

キャッシュメモリ107



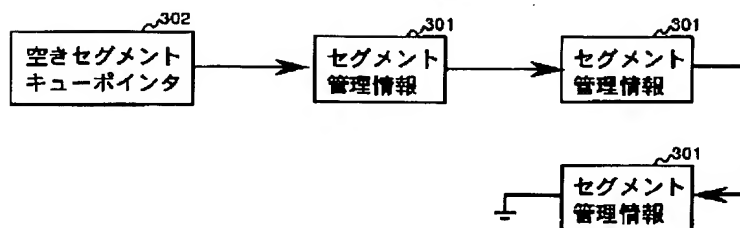
【図1】

図1



【図6】

図6



【図3】

図3

キャッシュディレクトリ 108

キャッシュ割り当て情報	300
セグメント管理情報	301
⋮	
セグメント管理情報	301
空きセグメントキューポインタ	302

【図5】

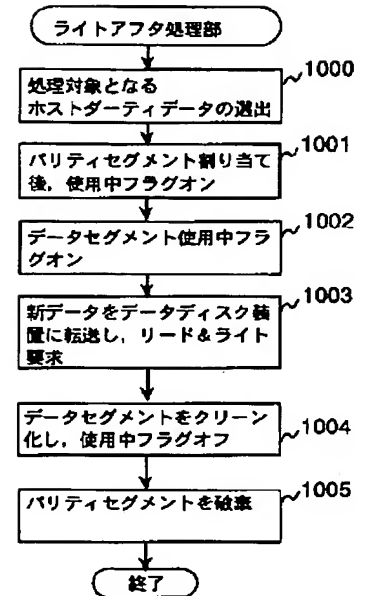
図5

セグメント管理情報 301

ライト面セグメントポインタ	500
リード面セグメントポインタ	501
ホストダーティフラグ	502
物理ダーティフラグ	503
パリティ未反映フラグ	504
ライト面バリッドフラグ	505
リード面バリッドフラグ	506
使用中フラグ	507
レコード状態	508
キューポインタ	509

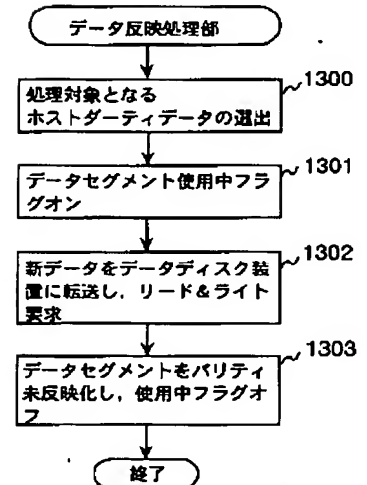
【図 10】

10



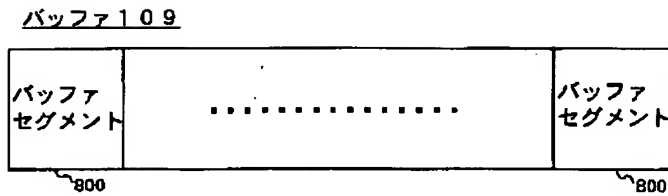
【図 13】

图 13



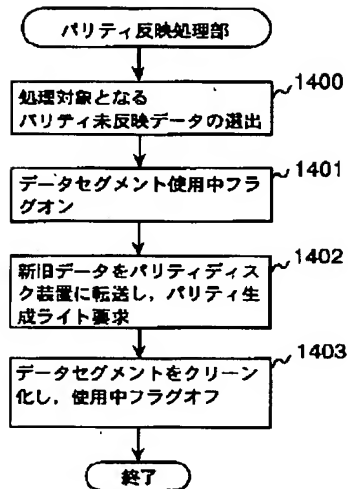
【図8】

図8



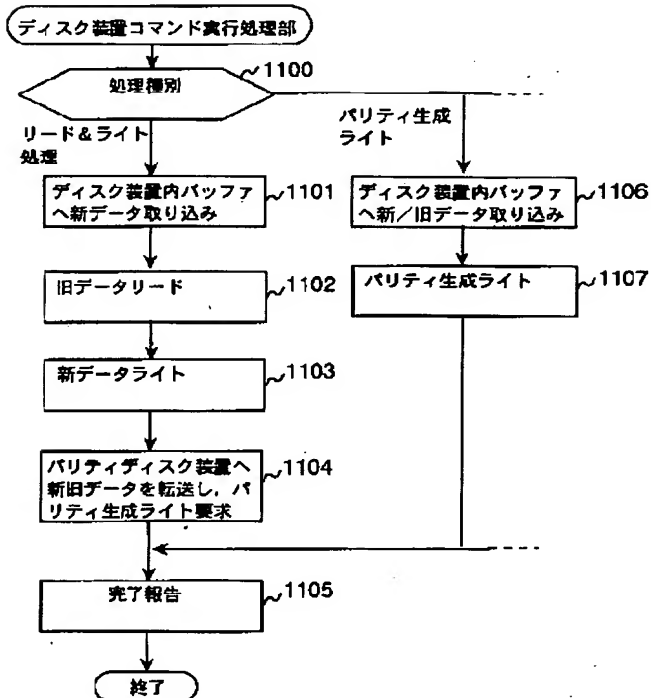
【図14】

図14



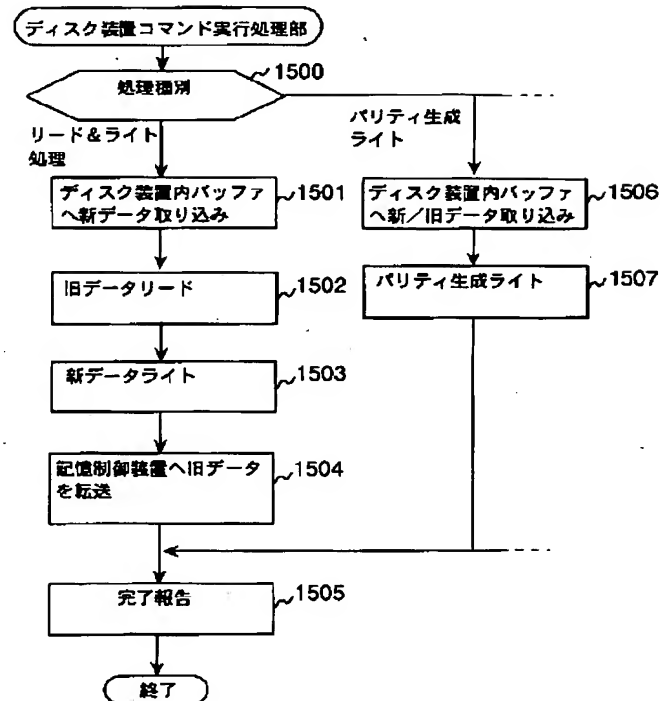
【図11】

図11



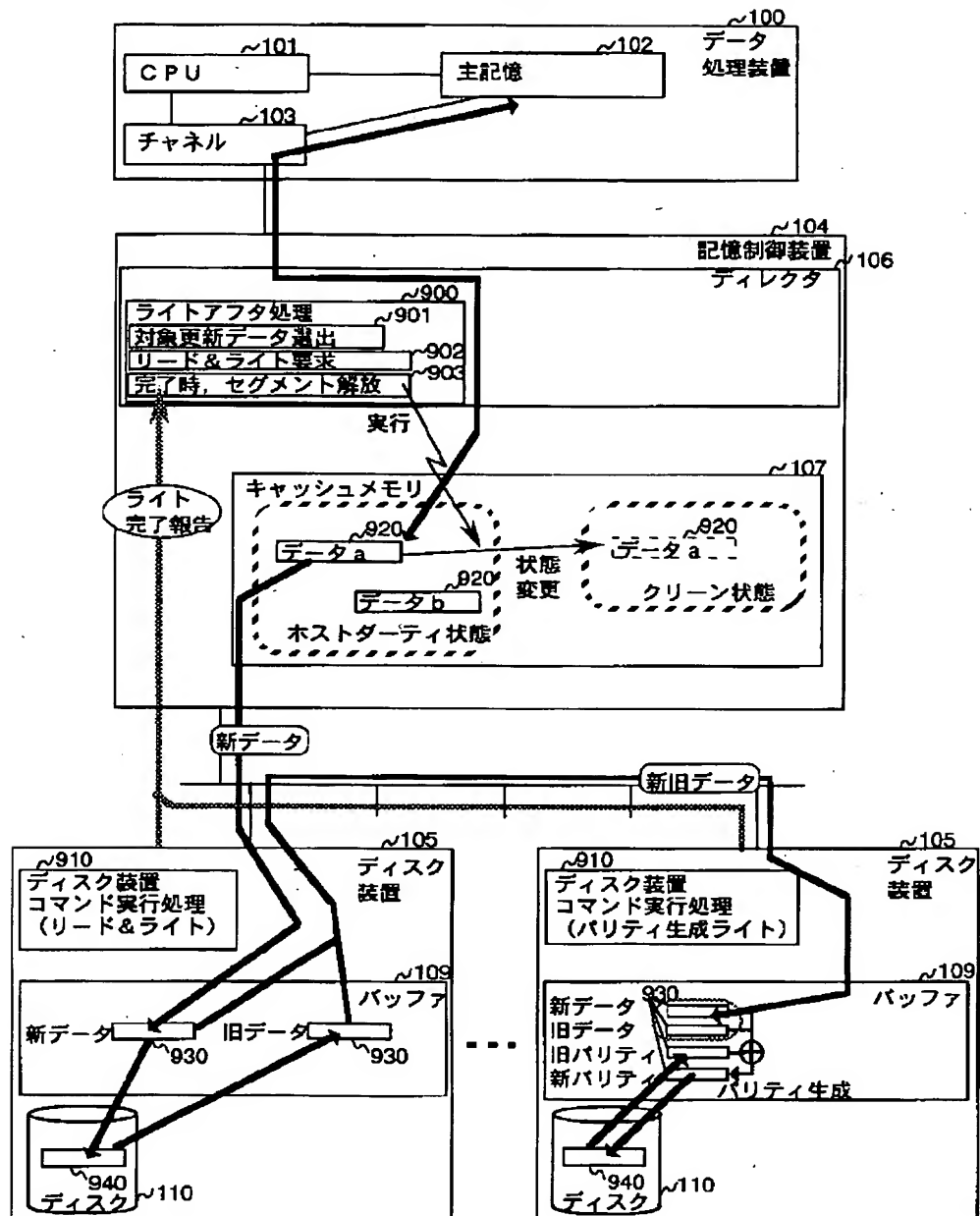
【図15】

図15



【図9】

図9



【図12】

図12

